Chapter Three
Descriptive Statistics

Objective:

1. Distinguish between measures of central tendency, measures of variability, and measures of shape.
2. Understand conceptually the meanings of mean, median, mode, quartile, percentile, and range.
3. Compute mean, median, mode, percentile, quartile, range, variance, standard deviation, and mean absolute deviation etc...
4. Understand the meaning of standard deviation as it is applied using the **empirical rule** and **Chebyshev's theorem**.

7. Understand box and whisker plots, skewness, and kurtosis.

<u>Descriptive statistics</u>:
It used to characterize and summarize quantitative variables. These descriptive statistics fall into basically three types of measurements: a) measures of center and b) measures of dispersion and c) measure of position. The following explains some of them:

Measures of Location:

1. Measure of Central Tendency.
   a. Population / Sample Mean
   b. Median
   c. Mode
2. Other Measures of location.
   d. Weighted Mean → population/ sample
   e. Percentiles
   f. Quartiles

Measures of Variation:

1. Range
2. Interquartile Range
3. Population Variance
4. Sample Variance
5. Population Deviation
6. Sample Deviation

Mean and standard deviation combined

1. Coefficient of Variation for Population
2. Coefficient of Variation for Sample
3. Empirical rule
4. Chebyshev's theorem
5. Standardized Data Value

3.1 Measures of Center

Central tendency:

For a set of data, we determine a quantity used to summarise the whole set of data. This quantity is termed a measure of central tendency. The most commonly used measures are **mean**, **medium** and **mode**.

Mean:

The mean is the sum of the observations divided by the number of observations (**Average** of a numerical set of data). We use $\overline{X}$ (x-bar) to denote sample mean and (μ) to denote population mean. It's often called the arithmetic mean.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad\qquad \mu = \frac{\sum_{i=1}^{n} x_i}{N}$$

$\sum$ is the Greek symbol sigma denotes the summation of all x values.
x is the variable usually used to represent the individual data values
n represents the number of data values in a sample
N represents the number of data values in a population

For ungrouped data: the following equation was used:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{it means that:} \quad \bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Example (3.1) the data represent the number of textbooks purchased by a sample of seven students:

(10, 4, 7, 5, 7, 8, 9)

Solution: $\overline{X} = \frac{10+4+7+5+7+8+9}{7} = 7.14$

For grouped data (mean in frequency distribution table):

a. Class mark method: Let $x_1, x_2, \ldots, x_n$ be a class marks and $f_1, f_2, \ldots, f_n$ be its frequencies respectively, where n is number of classes data. Arithmetic mean is given by:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i\ f_i}{\sum_{i=1}^{n} f_i} \quad \text{it means that:} \quad \bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \cdots + x_n f_n}{\sum_{i=1}^{n} f_i}$$

b. Assumption mean method: Let $x_1, x_2, \ldots, x_n$ be a class marks and $f_1, f_2, \ldots, f_n$ be its frequencies respectively, where n is number of classes data, and **(A) is Assumption mean which is the class mark with largest frequency** then the Arithmetic mean is given by:

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i\ (x_i - A)}{\sum_{i=1}^{n} f_i}$$

Example (3.2) the following frequency distribution gives the monthly consumption of electricity of 80 consumer of a locality. Find the mean of the data:

Class limit	31-40	41-50	51-60	61-70	71-80	81-90	91-100
Frequency	1	2	5	15	25	20	12

Solution:

Class limit	f_i	class mark method		assumption mean method	
		Class mark (x_i)	$x_i\ f_i$	x_i -A	$f_i\ (x_i - A)$
31-40	1	35.5	35.5	-40	-40
41-50	2	45.5	91	-30	-60
51-60	5	55.5	277.5	-20	-100
61-70	15	65.5	982.5	-10	-150
71-80	25	75.5	1887.5	0	0
81-90	20	85.5	1710	10	200
91-100	12	95.5	1146	20	240
Sum	80		6130		90

A = 75.5

Mean (class mark method) = (6130/80) = 76.625

Mean (assumption mean method) = 75.5+ (90/80) = 76.625

Weighted Arithmetic mean:

let $x_1,x_2,\dots,x_n$ be a class marks and $f_1,f_2,\dots,f_n$ be its frequencies respectively, where n is number of classes data then the Weighted Arithmetic mean is given by:-

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i\, w_i\, f_i}{\sum_{i=1}^{n} w_i\, f_i}$$

Example (3.3)

Class Limit	Frequency (fi)	Weighted (wi)	Class Mark (xi)	wi fi	wi fi xi
0-4	2	6	2	12	24
5-9	12	5	7	60	420
10-14	35	3	12	105	1260
15-19	38	5	17	190	3230
20-24	11	4	22	44	968
25-29	2	3	27	6	162
$\sum$				**417**	**6064**

Mean = 14.54

Median:

The median is a measure of central tendency more resistant to the effects of extreme values. The median is the value that occupies the middle position of data when data are put in **rank order by magnitude** (low to high OR high to low).

For ungrouped data:

Let n be the number of cases in your data. If n is odd, the median is the middle number of the data values sorted by magnitude. It occupies the $\left(\frac{n+1}{2}\right)^{th}$ position.

If n is even, the median is the average of the middle two numbers of the data sorted by magnitude. It is the average of the numbers in the $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n+2}{2}\right)^{th}$ positions.

For (odd number of values): find median of (1 3 **4** 8 10)?
The middle value is 4 (two values are higher, and two lower. This is the median.
While; for (even number of values): find median of (2 3 4 **4** **5** 8 9 9)?
The two middle values are 4 and 5. The median is the average of these two values (4+5)/2 = 4.5

* When the data follows a discrete set of values grouped by size, we use the formula $((n+1)/2)^{th}$ item for finding the median. First we form a cumulative frequency distribution, and the median is that value which corresponds to the **cumulative frequency** in which $((n+1)/2)^{th}$ item lies.

Median in Frequency Table

Example (3.4): find median for the following data

marks	1	2	3	4	5	6	7
No. students	2	11	15	20	25	18	10

Solution:

x_i	f_i	CF
1	2	2
2	11	13
3	15	28
4	20	48
5	25	73
6	18	91
7	10	101
∑	**101**	

Solution: n=101 (odd value),

Use median as size of $\left(\frac{n+1}{2}\right)^{th}$ position.

Median= size of (101+1) /2 = 51^{th} (greater and nearest value in cumulative frequency column)
Median = 5 (because 51^{th} item corresponds to 5).

If The Class Limit are Available:

first we find the **cumulative frequencies** and then calculate the **median class** by:

- Find $\left(\sum_{i=1}^{n} f_i\right)/2$.
- In cumulative frequency column; find the value greater than and nearest to (n/2). This row will be the median class.
- Find the median by applying this formula: $$\boxed{M_e = L + \left(\frac{\frac{\sum fi}{2} - cf}{f}\right) * h}$$

Where;
L = lower limit of the median class. (i.e. the class for which the cumulative frequency is just in excess of n/2).
h = length of the interval class.
n $=\Sigma f_i$ is the total number of observations
cf = cumulative frequency for the class **preceding** the median class.
f = frequency of the median class.

Example (3.5): find median for the data in previous example (example 3.2).

Class limit	*xi (class mark)*	*f_i*	Cum. Freq. *(cf)*
31-40	35.5	1	1
41-50	45.5	2	3
51-60	55.5	5	8
61-70	65.5	15	23
71-80	75.5	25	48
81-90	85.5	20	68
91-100	95.5	12	80
$\sum$		80	

Solution: $\sum f_i = 80$ (even value)
Position = $(\sum f_i /2)^{th} = (80/2)^{th} = (40)^{th}$
48 is the nearest greater value from 40
Therefore; 71-80 is the median class
L= 71, cf = 23, f= 25, h=9

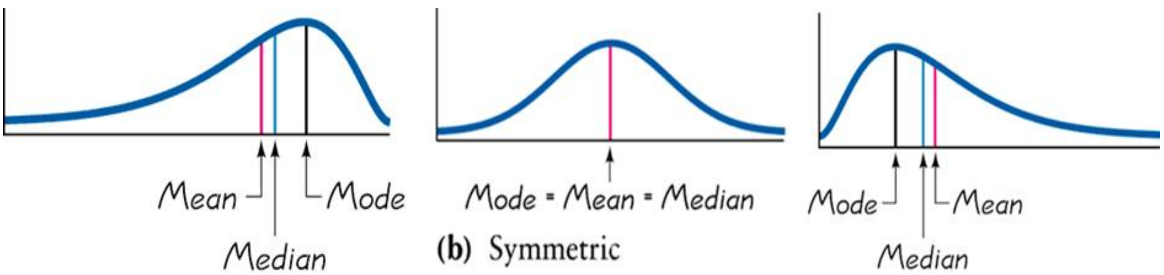
$$M_e = L + \left(\frac{\frac{\sum fi}{2} - cf}{f}\right) * h$$

$$M_e = 71 + \left(\frac{40-23}{25}\right) * 9 \;>>>\; M_e = 77.12$$

Shape of Distribution:

Sometimes mean, median and mode may not be able to reflect the true picture of some data. The following example explains the reason.

- **Symmetric:** Distribution of data is symmetric if the left half of its histogram is roughly a mirror image of its right half (have means and median that are basically the same**).**
- **Skewed:** Distribution of data is skewed if it is not symmetric and extends more to one side than the other.
- **Skewed to the left:** Also called negatively skewed have a longer left tail, mean and median are to the left of the mode (have a mean larger than median because the extreme value is below and pulling to the distribution to the left.
- **Skewed to the right:** Also called positively skewed have a longer right tail, mean and median are to the right of the mode (have a mean larger than median).



(a) Skewed to the Left (Negatively)
(b) Symmetric
(c) Skewed to the Right (Positively)

Mode

The most frequency value of data is the mode and denoted by M_o. this is the only measure of center that can be used with qualitative data.

Example (3.6):

Find the mode of:

a. 12, 12,13,12,15,17 M_o=12
b. 3,3,3,7,7,7,8,8,8,10,10 M_o= 3,7,8
c. 9,12,300 No Mode
d. 7,0,0,8,0,2,0,5,0 M_o=0

Types of mode:

Unimodal distribution: a histogram with one very high area.

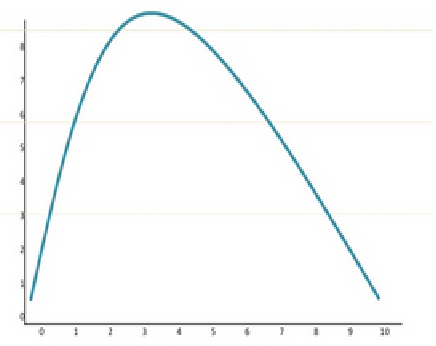
Bimodal distribution: has two peaks. Two peaks in a bimodal distribution also represent two local maximums; these are points where the data points stop increasing and start decreasing.

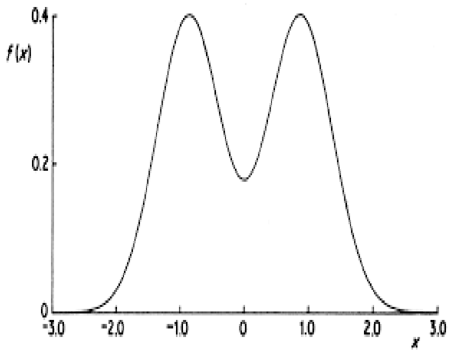
Multimodal distribution: is a probability distribution with more than one peak, or "mode". If you can't clearly find one peak or two peaks in a graph, the likelihood is that you either have a uniform distribution (where all the peaks are the same height) or a multimodal distribution, where there are several peaks of the same height.

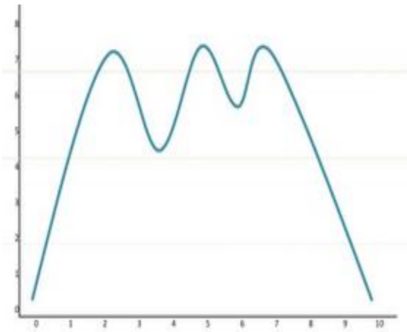
No Mode: When no data value is repeated, we say that there is no mode.

Unimodal Mode

Multimodal Mode Bimodal Mode







NOTE:

- **If all are of the same frequency, no mode exits. (When no value is repeated).**
- **If more than one value has the same largest frequency, then the mode is not unique (There can be more than one mode).**

Example (3.7):

Marks (xi)	42	36	30	45	50
No. students (fi)	7	10	**13**	8	2

Solution:

Mode= 30 (because it's corresponding to the largest frequency which was 13).

Mode in Frequency Table:

Mode can be found by first identify the **largest frequency** of that class, called **modal class**, and then apply the following formula:

$$M_o = L + \left(\frac{d_1 - d_0}{2d_1 - d_0 - d_2}\right) * h$$

Where;
L is the lower limit of the modal class, h is the length of interval class, d_o is the frequency of the class preceding the modal class, d_1 is the frequency of the modal class, d_2 is the frequency of the class succeeding the modal class.

Example (3.8):

Find the mode of the grouped data:

Class limit	5-10	20-35	35-50	50-65	65-80
Frequency	3	4	8	9	6

Solution:

Modal class = (50-65), L= 50, h=15, d_o=8, d_1=9 and d_2=6

Apply mode formula; M_o = 53.75

Example (3.9): (homework)

The data in Table below (Adamson, 1989) are the annual maximum flood peak flows to the Hardap Dam in Namibia with catchment area 12600 km^2, covering the period from October1962 to September 1987. The range of these data is from 30 to 6100. Calculate Mean, Median and Mode?

Year	1962-1963	1963-1964	1964-1965	1965-1966	1966-1967	1967-1968	1968-1969	1969-1970	1970-1971
Inflow (m^3/s)	1864	44	46	364	911	83	477	457	782
Year	1971-1972	1972-1973	1973-1974	1974-1975	1975-1976	1976-1977	1977-1978	1978-1979	1979-1980
Inflow (m^3/s)	**6100**	197	3259	554	1506	1508	236	635	230
Year	1980-1981	1981-1982	1982-1983	1983-1984	1984-1985	1985-1686	1986-1987		
Inflow (m^3/s)	125	131	**30**	765	408	347	412		

Solution: The inflow data are used to construct the frequency distribution table.

No. of data =	25		
No. of classes =	5.643957	take	6
class width =	1011.667	take	1012

No. of classes	class limit		fi	class mark	fi * xi	class boundary		Relative frequency %	<CF	Class Name
	lower	upper				lower	upper			
1	30	1041	20	535.5	10710	29.5	1041.5	80	20	Modal, Median
2	1042	2053	3	1547.5	2642.5	1041.5	2053.5	12	23	
3	2054	3065	0	2559.5	0	2053.5	3065.5	0	23	
4	3066	4077	1	3571.5	3571.5	3065.5	4077.5	4	24	
5	4078	5089	0	4583.5	0	4077.5	5089.5	0	24	
6	5090	6101	1	5595.5	5595.5	5089.5	6101.5	4	25	
∑			25		24519.5					

Mean: $\bar{x}$ = 980.78

Median: M_e= 661.87

Mode: M_o = 576.48

Example (3.10): (Homework)

Unit weight measurements from a boring are presented in Table below. This boring was drilled offshore in the Gulf of Mexico at the location of an oil production platform. The soil consists of a normally consolidated clay over the length of the boring. The unit weight varies with depth, and ranges from 95 to 125 Ib/ft^3.

Required:

1- Determine Mean, Median, and Mode from the Frequency Table.

2- Draw an Ogive chart between Depth and Total Unit weight, and from the graph.

Total Unit Weight Data from Offshore Boring								
Depth (ft)	0.5	1.0	1.5	5.0	6.5	7.5	16.5	19.0
Total Unit Weight, (Ib/ft^3)	105	119	117	99	101	96	114	100
Depth (ft)	22.0	25.0	27.5	31.0	34.5	37.5	40.0	45.0
Total Unit Weight, (Ib/ft^3)	99	102	100	101	101	100	101	99
Depth (ft)	50	60.5	62.0	71.5	72.0	81.5	82.0	91.5
Total Unit Weight, (Ib/ft^3)	100	103	101	106	109	100	104	102
Depth (ft)	101.5	102.0	112.0	121.5	122.0	132.0	142.5	152.5
Total Unit Weight, (Ib/ft^3)	106	99	102	100	101	101	104	102
Depth (ft)	162.0	172.0	191.5	201.5	211.5	241.5	251.5	261.8
Total Unit Weight, (Ib/ft^3)	105	**95**	116	107	112	114	109	110
Depth (ft)	271.5	272.0	281.5	292.0	301.5	311.5	322.0	331.5
Total Unit Weight, (Ib/ft^3)	109	106	108	111	**125**	112	104	113
Depth (ft)	341.5	342.0	352.0	361.5	362.0	371.5	381.5	391.5
Total Unit Weight, (Ib/ft^3)	112	113	116	124	117	114	115	114
Depth (ft)	392.0	402.0	411.5	412.0	421.5	432.0	442.0	451.5
Total Unit Weight, (Ib/ft^3)	115	114	112	115	115	112	115	119

Solution:

No. of data	64		
No. of classes	7.00013	take	7
class width	4.285714	take	5

No. of classes	Class limits		Frequency (f_i)	(<CF)	Class Mark (x_i)	Class Boundary		$f_i\, x_i$	Class Name
	Lower	Upper				Lower	Upper		
1	95	99	6	6	97	94.5	99.5	582	
2	100	104	**21**	21+6=27	102	99.5	104.5	2142	Modal
3	105	109	10	27+10=**37**	107	104.5	109.5	1070	Median
4	110	114	14	37+14=51	112	109.5	114.5	1568	
5	115	119	11	51+11=62	117	114.5	119.5	1287	
6	120	124	1	62+1=63	122	119.5	124.5	122	
7	125	129	1	63+1=64	127	124.5	129.5	127	
$\sum$			64					6898	

Mean: $\bar{x} = 6898/64 = 107.78$

Median: $M_e = 128.6$

Mode: $M_o = 102.3$